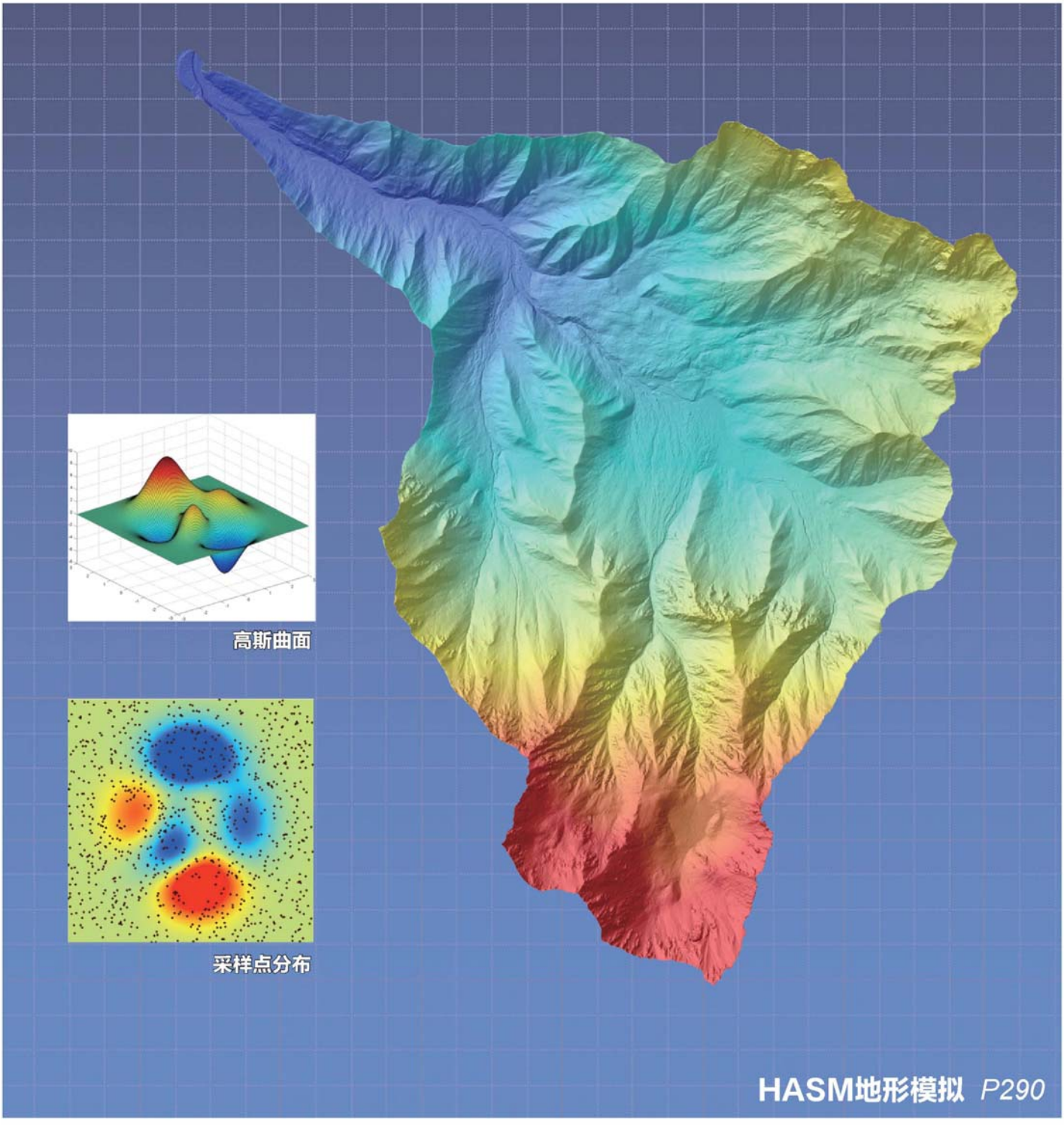
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2014
02
VOL.19

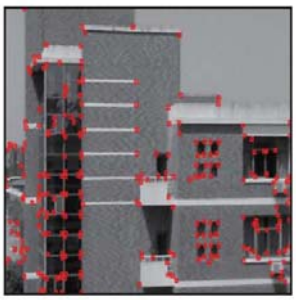
ISSN1006-8961
CN11-3758/TB



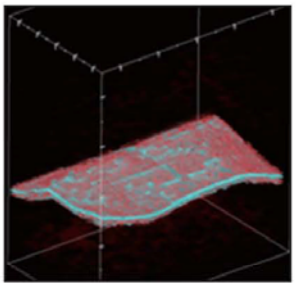
2014年2月 第19卷 第2期（总第214期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

邻域最短距离法寻找最佳拼接缝(第227页)



曲率尺度空间与链码方向统计的角点检测(第234页)

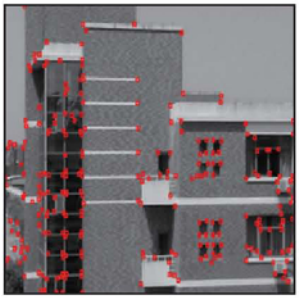


目标特征指导的多分辨率体绘制算法(第283页)

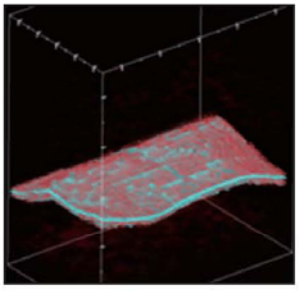
Volume 19, Number 2
Published February 16, 2014

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Finding an optimal seam-line through the shortest distance in the neighborhood(P227)



Corner detection based on curvature scale space and chain code direction statistics(P234)



Target feature based multi-resolution volume rendering(P283)

E-mail: jig@irsa.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号： **文献标识码**：A **文章编号**：1006-8961(2014)02-0175-10

论文引用格式：郑胤,陈权崎,章毓晋. 深度学习及其在目标和行为识别中的新进展[J]. 中国图象图形学报,2014,19(2):175-184. [DOI: 10.11834/jig.20140202]

深度学习及其在目标和行为识别中的新进展

郑胤，陈权崎，章毓晋

清华大学电子工程系，北京 100084

摘 要：**目的** 深度学习是机器学习中的一个新的研究领域。通过深度学习的方法构建深度网络来抽取特征是目前目标和行为识别中得到关注的研究方向。为引起更多计算机视觉领域研究者对深度学习进行探索和讨论,并推动目标和行为识别的研究,对深度学习及其在目标和行为识别中的新进展给予概述。**方法** 首先介绍深度学习领域研究的基本状况、主要概念和原理;然后介绍近期利用深度学习在目标和行为识别应用中的一些新进展。**结果** 阐述了深度学习与神经网络之间的关系,深度学习的优缺点,以及目前深度学习理论需要解决的主要问题。**结论** 该文对拟将深度学习应用于目标和行为识别的研究人员有所帮助。

关键词：深度学习;目标识别;行为识别;计算机视觉

Deep learning and its new progress in object and behavior recognition

Zheng Yin, Chen Quanqi, Zhang Yujin

Department of Electronic Engineering, Tsinghua University, Beijing 100084, *China*

Abstract: **Objective** Deep learning is a new research area in machine learning. Currently, extracting features by deep learning for visual object recognition and behavior recognition capture many attentions. To draw more attention from research community about deep learning, and to push forward the research frontier of object and behavior recognition, we give a general progress overview for deep learning and its application to visual object and behavior recognition. **Method** First, we give a general introduction to deep learning, including the basic situation, main concepts and principle. Then, some new progresses on using deep learning in visual object recognition and behavior recognition are presented. **Result** A discussion about the differences between deep learning and neural network as well as the advantage and disadvantage of deep learning are given, the main existing problems that should be solved for deep learning theory are pointed. **Conclusion** This paper should provide some help for the research community on applying the deep learning to the visual object and behavior recognition.

Key words: deep learning; object recognition; behavior recognition; computer vision

0 引 言

计算机视觉是指用计算机实现人的视觉功能,希望能根据感知到的图像(视频)对实际的目标和场景内容做出有意义的判断[1]。如何能正确识别目标和行为非常关键,其中一个最基本的和最核心的问题是对图像的有效表达。如果所选的表达特征能够有效地反映目标和行为的本质,那么对于理解图像就会取得事半功倍的效果。正因为如此,关于

收稿日期:2013-06-28;**修回日期**:2013-11-18

基金项目:国家自然科学基金项目(61171118);教育部高等学校博士学科点专项科研基金项目(SRFDP-20110002110057)

第一作者简介:郑胤(1986—),男,清华大学电子工程系博士研究生,主要研究方向为深度学习、机器学习、模式识别、计算机视觉、图像工程。E-mail:y-zheng09@mails.tsinghua.edu.cn

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 2, Feb. 2014

特征的构建和选取一直得到广泛关注。近些年来人们已构建出许多特征,并且得到了广泛的应用,例如SIFT[2]、HOG[3]、LBP[4]、MSER[5]等等。设计特征是一种利用人类的智慧和先验知识,并且将这些知识应用到目标和行为识别技术中的很好的方式。但是,如果能通过无监督的方式让机器自动地从样本中学习到表征这些样本的更加本质的特征则会使得人们更好地用计算机来实现人的视觉功能,因此也是近些年人们关注的一个热点方向。深度学习(deep learning)的目的就是通过逐层的构建一个多层的网络来使得机器能自动地学习到反映隐含在数据内部的关系,从而使得学习到的特征更具有推广性和表达力。

本文旨在向读者介绍深度学习的原理及它在目标和行为识别中的最新动态,希望吸引更多的研究者进行讨论,并在这一新兴的具有潜力的视觉领域做出更好的成果。首先对深度学习的动机、历史以及应用进行了概括说明;主要介绍了基于限制玻尔兹曼机(RBM)[6-7]的深度学习架构和基于自编码器(auto-encoder)[8-9]的深度学习架构,以及深度学习近些年的进展,主要讨论了去噪自编码器(denoising autoencoder)[10],卷积限制玻尔兹曼机(convolutional RBM)[11],三元因子玻尔兹曼机(3-way factorized Boltzmann machine)[12-13],以及神经自回归分布估计器(NADE)等一些新的深度学习单元;对目前深度学习在计算机视觉中的一些应用以及取得的成果进行介绍;最后,对深度学习与神经网络的关系,深度学习的本质等问题加以讨论,提出目前深度学习理论方面需要解决的主要问题。

1 深度学习概述

目前在典型使用的技术中是通过"特征表达"+"分类器"的框架来进行目标识别、行为识别等任务的,如图1所示。

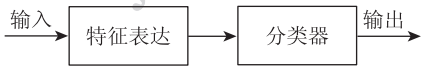


图1 计算机视觉用于识别的框架

Fig. 1 The framework of recognition in computer vision

传统的"特征表达"是通过人们手动设计的特征提取到的,也就是说在目前的计算机视觉框架内存在一个对输入信号的一个"显式"的预处理过程。但是最近神经科学关于哺乳动物的信息表达的研究发现[14-15],哺乳动物大脑中关于执行识别任务的大脑皮层并没有一个"显式"的对信号预处理的过程,而是将输入信号在一个大脑的复杂的层次结构中传播,通过每一层次对输入信号进行重新的提取和表达最终让哺乳动物感知世界。这些研究促成了深度学习这一机器学习子领域的兴起[16],它试图通过让计算机模拟人脑感知视觉信号的机制,进而设计深层的网络来实现视觉的功能。目前深度学习已经成为计算机视觉中的一个热点方向,每年都有大量的研究成果出现,产生了诸多深度学习的新算法和新方向,而同时深度学习算法的性能也逐渐在一些国际重大评测中超过了其他方法[17-18]。

2 深度学习原理

传统的随机初始化模型参数然后用反向传播(back-propagation)来优化参数的方法对于深度网络来说容易造成陷入局部极值或者产生梯度弥散等问题,因此人们提出使用额外的目标函数来对每层的参数进行预处理,然后对预处理之后的模型进行反向传播来进一步优化参数。这其中限制玻尔兹曼机和自编码器是两个常用的预处理单元,而基于这两个单元的深度模型也成为了当前深度学习的主流框架。深度学习的算法体系如图2所示,根据学习单元的不同,深度学习主要包括基于限制玻尔兹曼机的深度置信度网络(DBN)[6]和基于自编码器的深度网络(stacked auto-encoder)[8-9]两类,另外还有一些其他体系的深度网络。本节主要介绍上面两种主流的深度

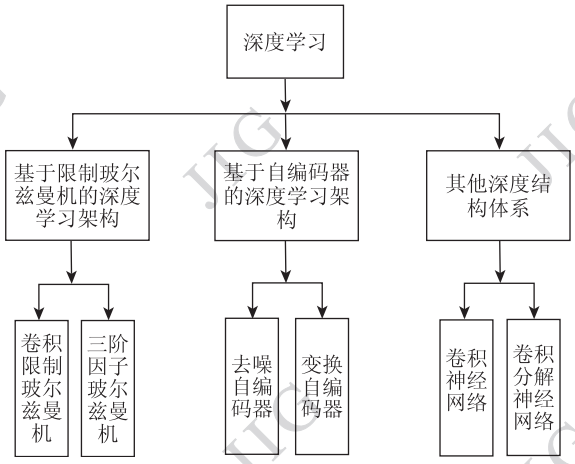


图2 目前深度学习算法体系结构

Fig. 2 The family of deep learning algorithm

学习架构的原理以及在实际操纵中经常要用到的稀疏性约束的原理和做法。

2.1　基于限制玻尔兹曼机的深度学习架构

限制玻尔兹曼机是构成深度置信网络的基础单元，其本质是使得学习到的模型产生符合条件的样本的概率最大。

2.1.1　玻尔兹曼机

玻尔兹曼机(Boltzmann machine)[19]本质上是一种能量模型。能量模型是指对于参数空间(configuration space)中每一种情况均有一个标量形式的能量与之对应。能量函数就是从参数空间到能量的映射函数，人们希望通过学习使得能量函数有符合要求的性质。从结构上来说，玻尔兹曼机是双层、无向、全连通图，如图 3 所示。为了方便起见，这里仅讨论观测变量和隐变量均是 0、1 变量的情况。

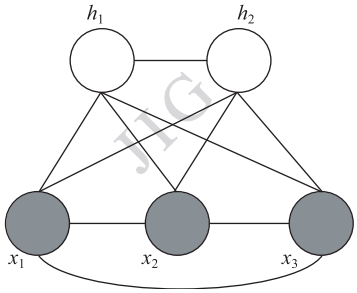


图 3　玻尔兹曼机示意图

Fig. 3　The illustration of Boltzmann machine

玻尔兹曼机的能量函数为

$$E(x,h) = -b'x - c'h - h'\boldsymbol{W}x - x'\boldsymbol{U}x - h'\boldsymbol{V}h \quad (1)$$

式中，x 表示可见层，h 表示隐层，$b \in \{0,1\}^K$，$c \in \{0,1\}^D$ 分别表示可见层和隐层单元的偏置(offset)，K、D 分别表示可见层和隐层单元的数目。$\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{V}$ 分别表示观测层和隐层之间，观测层变量之间，隐层变量之间的连接权重矩阵。

在实际中，由于计算样本概率密度时归一化因子的存在，需要使用马尔可夫蒙特卡洛方法(MCMC)[20]来对玻尔兹曼机进行优化。但是 MCMC 方法收敛速度很慢，因此人们提出限制玻尔兹曼机和对比散度方法来解决这一问题。

2.1.2　限制玻尔兹曼机

限制玻尔兹曼机[21]是对全连通的玻尔兹曼机进行简化，其限制条件是在给定可见层或者隐层中的其中一层后，另一层的单元彼此独立，即式(1)中 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 矩阵中的元素均等于 0。层间单元独立的条件是构成高效的训练限制玻尔兹曼机的方法的条件之一[6]，而 RBM 也因此成为深度置信网络(DBN)[6]的构成单元。限制玻尔兹曼机的图模型如图 4 所示。可见，层内单元之间没有连接关系，层间单元是全连接关系。

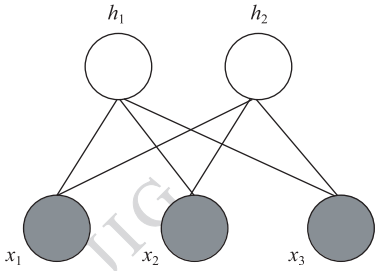


图 4　限制玻尔兹曼机示意图

Fig. 4　The illustration of restricted Boltzmann machine

将式(1)中层间连接矩阵 $\boldsymbol{U}$，$\boldsymbol{V}$ 置零，得到限制玻尔兹曼机的能量函数

$$E(x,h) = -b'x - c'h - h'\boldsymbol{W}x \quad (2)$$

由于限制玻尔兹曼机取消了层内单元之间的连接，所以可以将其条件概率分布进行分解，这样就简化了模型优化过程中的运算。但是在其优化过程中仍然需要基于 MCMC 方法的吉布斯采样，训练过程仍然十分漫长，因此人们提出对比散度方法来加快模型优化。

2.1.3　对比散度

对比散度(contrastive divergence)是 Hinton[6]在 2006 年提出来的快速地训练限制玻尔兹曼机的方法，该方法在实践中得到广泛的应用。对比散度主要是将对数似然函数梯度的求解进行了两个近似：

1)使用从条件分布中得到的样本来近似替代计算梯度时的平均求和。

这是因为在进行随机梯度下降法进行参数优化时已经有平均的效果，而如果每次计算都进行均值求和则这些效果会相互抵消，而且会造成很大的计算时间的浪费。

2)在进行吉布斯采样(Gibbs sampling)时只采用一步，即仅仅进行一次吉布斯采样。

这种一次吉布斯采样方法会使得采样得到的样本分布与真实分布存在一定的误差。但是实践发现，如果仅作一次迭代的话，就已经能得到令人满意的结果。

将限制玻尔兹曼机逐层叠加，就构成了深度置信网络(DBN)。在深度置信网络中底层的输出作

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 2, Feb. 2014

为上一层的输入,每层是一个限制玻尔兹曼机,使用对比散度的方法单独训练。为了达到更好的识别效果,往往还要对深度置信网络每层的参数进行微调[6,22]。使用限制玻尔兹曼机构建成深度网络,在一些公开的数据集上取得了非常好的效果[23]。

2.2 基于自编码器的深度学习架构

另一种主流的构成深度学习架构的单元是自编码器[8-9],其每一层学习单元的目的是使得重建误差最小。自编码器的示意图如图5所示。自编码器的核心思想是将输入信号进行编码,使用编码之后的信号重建原始信号,目的是让重建信号与原始信号相比重建误差最小。自编码器的思想在计算机视觉中有广泛的应用,通过将信号编码成为另一种形式,可以有效地提取信号中的主要信息,去除冗余,并且能够更加简洁地表达。从某种意义上来说,可以将计算机视觉中经常用到的 K 均值聚类、稀疏编码、主成分分析等方法均理解为是一个自编码器。

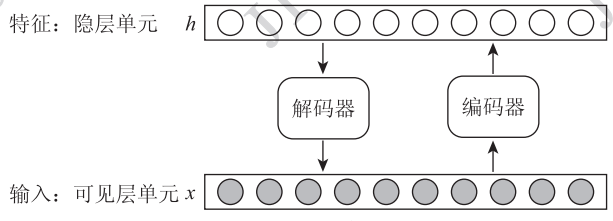


图5 自编码器示意图

Fig. 5 The illustration of auto-encoder

如果在编码和解码中使用线性函数,误差函数是均方误差,则这种自编码器就等价于主成分分析;而如果使用量化编码,误差函数是均方误差,则这种自编码器等效于 K 均值聚类、稀疏编码等。

由于自编码器的编解码过程以及目标函数都是确定性的,因此不必像限制玻尔兹曼机一样采用马尔可夫蒙特卡洛的方法作近似,所以它的优化过程仅需要使用根据目标函数对于各个参数的导数采用随机梯度下降法即可。

将自编码器逐层叠加就构成了深度自编码器(stacked auto-encoder)。深度自编码器的下一层的输出作为上一层的输入,每一层单独进行优化。这样每一层都可以对原始信号作编码表达,这样最终得到隐含在信号内部的深层的数据关系,因此可以作为信号的更加本质和有效的表达。最后,为了充分利用数据中的类别信息,还需要使用监督的方法对参数进行微调。这种微调一般通过在顶层增加一个逻辑回归层(logistic regression layer)来实现。

2.3 稀疏性约束

深度置信网络和深度自编码器都是通过逐层学习来构建深度网络。由于经过深度网络处理之后的信号的维数一般远远高于原始信号,而往往这种方式学习到的关于原始信号的表达不是稀疏的。而在计算机视觉中,稀疏性的约束是使得学习到的表达更有意义的一种重要约束,并且在独立成分分析(ICA[24]),稀疏编码(sparse coding[25])等算法中得到了验证。另一方面,由于深度学习中要优化的参数非常多,如果不加入稀疏性的约束往往会使得学习到的权重矩阵为单位矩阵,这样就失去了深度的意义。同时,神经科学中也有研究表明,人脑在感知视觉信号的时候各个神经元的响应也是稀疏的[25-26]。因此,人们希望在深度学习的优化过程中加入稀疏约束。

引入稀疏性约束的方法有两种:一是使隐层的响应稀疏,二是使隐层和可见层之间的连接权重矩阵稀疏。这两种稀疏性的约束目前为止还没有明确的优缺点分析,往往需要在实践当中结合具体的问题进行试验。为了简化起见,本节只介绍基于限制玻尔兹曼机的稀疏性约束,而基于自编码器的深度网络的稀疏性约束可以类似地得到。

2.3.1 隐层响应稀疏约束

由于人脑在感知信号的时候每次只有一部分神经元被激活[25-26],受此启发,这种稀疏性约束直接对隐层的响应 h 建模。其核心思想是:因为隐层是二值函数,因此使隐层单元的响应稀疏则隐层的所有单元的响应值的期望会比较小,因此 Lee 等人[27]提出在对样本出现概率进行优化的基础上,增加对每个隐层单元响应的约束,使得隐层单元的响应的 L_1 范数很小。因为每个隐层单元响应值介于0和1之间,这样会使得隐层的响应稀疏,其目标函数为

$$\min_{W,c,b}\left\{-\sum_{l=1}^{m}\log\sum_{h}P(x^{(l)},h^{(l)})+\lambda\sum_{j=1}^{K}\left|p-\frac{1}{m}\sum_{l=1}^{m}E[h_j^{(l)}\mid x^{(l)}]\right|^2\right\} \tag{3}$$

式中,$p\in(0,1)$ 表示隐层单元的响应的期望,一般数值比较小,$\sum_{h}P(x^{(l)},h^{(l)})$ 表示第 l 个样本 $x^{(l)}$ 出现的概率,$h^{(l)}$ 表示隐层的响应,$E[h_j^{(l)}\mid x^{(l)}]$ 表示第 j 个隐层单元响应的期望,m 表示样本的数量。

2.3.2　连接矩阵稀疏约束

另一种稀疏性的约束是直接对连接权重矩阵 W 作约束。其施加约束的方法如下：如果某一次迭代使得权重矩阵 $\boldsymbol{W}$ 中的某一元素 W_{ij} 改变符号，那么将这个元素 W_{ij} 置零。这种方法的原理如下：希望权重矩阵 $\boldsymbol{W}$ 尽量稀疏，这需要使用 L_0 范数。但是解 L_0 范数是一个 NP-Hard 问题，所以人们使用 L_1 范数作近似，因为 L_1 范数可以被有效地解决同时在满足一定条件下可以很好地近似 L_0 范数。但是 L_1 范数是非平滑的，它不会令 $\boldsymbol{W}$ 稀疏，但是如果在某一次迭代过程中权重矩阵 $\boldsymbol{W}$ 中的某一元素 W_{ij} 改变符号，那么就说明 W_{ij} 足够小，则可以将其置零。

3　深度学习的进展

自从 2006 年 Hinton 等人提出训练限制玻尔兹曼机的有效算法之后，深度学习以其优异的性能成为人们研究热点。这期间产生出诸多优秀的算法，进一步推动深度学习领域的发展。下面主要介绍近些年深度学习在计算机视觉中产生的一些有代表性的新方法。

3.1　去噪自编码器

自编码器是以编码之后的重建误差作为优化的目标，使得编码之后的信号捕捉到原信号中的主要信息。但是实际信号往往含有噪声，使得学习到的自编码器的性能有所下降。为了使得深度网络对于噪声更加鲁棒，Vincent 等人[10]提出了去噪自编码器，并且可以将去噪自编码器替代自编码器来构建深度网络。

去噪自编码器的核心思想是在信号上施加噪声作为训练集，将重建信号与未施加噪声的信号作对比来作为重建误差，这样就使得自编码器具有一定的抗噪声能力。如图 6 所示，x 是原始信号，$\tilde{x}$ 是施加噪声之后的信号，y 是对 $\tilde{x}$ 进行编码之后的信号，z 是重建信号。而重建误差是用 x 和 z 来计算。通过优化重建误差，使得去噪自编码器能够适应一定的噪声干扰。去噪自编码器原理简单，但是在实践中非常有效，能够有效解决样本中的噪声干扰问题。

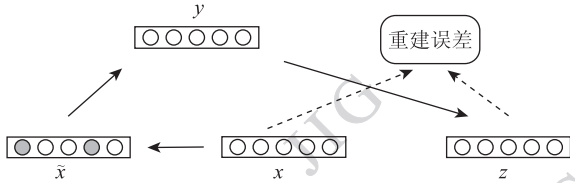


图 6　去噪自编码器示意图

Fig. 6　The illustration of denoising auto-encoder

3.2　三元因子玻尔兹曼机

传统的限制玻尔兹曼机要求层内单元之间无连接。这种假设使得可以将每层单元的条件概率彼此独立，因此可以有效地计算每层的条件概率。但是对于图像数据来说，像素之间是存在相互关系的，如果能将像素之间的关系考虑进来，并且能够有效地计算，则理论上可以更好地表达图像和理解图像。三元因子玻尔兹曼机[12-13]就是将像素之间的两两关系考虑进来，使得隐层能够捕捉到可见层单元之间的关系。三元因子玻尔兹曼机的核心思想是通过因子来作为可见层两两单元和隐层单元之间连接纽带。图 7 是三元因子玻尔兹曼机构成单元的示意图。

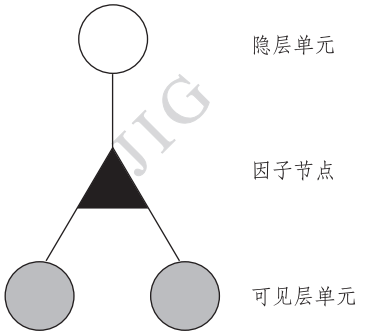


图 7　三元因子玻尔兹曼机构成单元示意图

Fig. 7　The illustration of a unit of factored 3-way Boltzmann machine

三元因子玻尔兹曼机的能量函数为

$$E(v,h) = -\sum_{i,j,k} v_i v_j h_k w_{i,j,k} \tag{4}$$

式中，v_i、v_j 表示第 i、j 个可见层单元，h_k 表示第 k 个隐层单元，$w_{i,j,k}$ 表示三者之间的连接权重。在实际中可以通过矩阵分解的方法对式(4)进行变形和优化。

Ranzato 等人在 CIFAR-10 图像数据库[13,28]上进行了实验，结果表明，引入可见层单元之间联系方式可以更好地对图像进行建模，引入可见层单元之间的两两关系，取得更好的识别效果。

3.3　卷积限制玻尔兹曼机

由于传统的深度学习算法将图像中的每一个像素作为一个可见单元，由于计算量的限制，这种处理方式往往只能处理比较小的图像(例如 32 × 32 像素的手写体字符)。但是在计算机视觉中人们希望能够对通常意义上的图像(100 × 100 像素以上)进行

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 2, Feb. 2014

处理,这样可见单元的数量会数以万计,即使使用GPU加速的方法也很难解决。另一方面,由于计算机视觉中要处理的图像大小不同,长宽比不同,也无法建立一个统一大小的连接权重矩阵。因此Lee等人[11]提出卷积限制玻尔兹曼机(Convolutional RBM)的算法来使得深度学习可以处理通常意义下的图像,并且可以自动地学习到目标的部件结构。

卷积限制玻尔兹曼机的基本思想是使用卷积的方式使得图像各个像素共享一组滤波器(类比于链接权重矩阵)$\boldsymbol{W}^k, k=1,2,\cdots,K$,通过$W^k$与图像像素作卷积的方式从图像中抽取特征,并且通过概率最大值汇总[11]的方式使得抽取的特征具有一定的平移不变性。如图8所示是卷积限制玻尔兹曼机的示意图,$\boldsymbol{V}$是输入的图像,即可见层,$\boldsymbol{H}$是隐层,$\boldsymbol{P}$是汇总层,$\boldsymbol{W}^k$是第k个滤波器,可见层的大小是$N_V \times N_V$,滤波器的大小为$N_W \times N_W$,这样参数的个数就从$O(N_V^2)$变成了$O(KN_W^2)$,而$N_W \ll N_V$,这样参数的个数就大大降低了,从而使得进行计算是可行的。

将卷积限制玻尔兹曼机逐层叠加,就得到深度模型,进而可以学习到不同层次的图像表达。卷积限制玻尔兹曼机可以应用到图像中的目标识别和分类,学习到属于不同层次的语义信息。

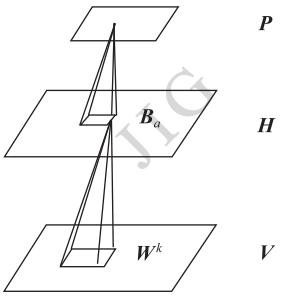


图8 卷积限制玻尔兹曼机示意图

Fig. 8 The illustration of convolutional RBM

3.4 神经自回归分布估计器

对于多维数据分布的估计是计算机视觉中处理数据时的一个基本问题。事实上,如果能够估计出样本中的分布,人们就能够回答关于数据的统计关系的任何问题,例如,给定某些观测量,推断出其他观测量出现的概率等等。从某种程度上来说,限制玻尔兹曼机可以被理解成为一种通过隐变量来对样本的分布作估计的分布估计器。通过训练限制玻尔兹曼机来学习观测到的数据的分布,可以得到给定观测数据后隐变量的分布并以此作为特征。当使用多层的限制玻尔兹曼机时,这种分布估计器其实就是深度网络。但是使用限制玻尔兹曼机在估计样本分布的时候需要计算归一化因子,这需要穷举所有可能的观测变量和隐变量的情况,因此只能对小规模隐变量的分布准确地估计。

为了解决限制玻尔兹曼机在估计多维数据的分布而计算归一化因子时遇到的问题,Larochelle等人[29]提出神经自回归分布估计器(NADE)。神经自回归分布估计器通过将观测变量的联合概率分布分解成一系列条件概率的乘积,并且使用类似于限制玻尔兹曼机的结构来对每一个条件概率进行建模,从而避免计算归一化因子而带来的计算复杂度方面的问题。其模型的结构如图9所示。图9中,W,b,c表示模型的参数,v_i表示输入变量,$\hat{v}_i$表示给定前$i-1$个输入的条件下,模型预测的第i个输入变量的出现概率。$\boldsymbol{h}_i$表示输入了$i-1$个输入后,隐层(只有一层)的响应。在文献[29]中,Larochelle等人使用随机梯度下降法对模型参数进行优化,并且证明该模型的计算复杂度为$O(HD)$,其中H为隐变量的个数,D为观测变量的维数。实验结果表明,神经自回归分布估计器能够比限制玻尔兹曼机更加有效地对输入样本的分布进行估计[29]。

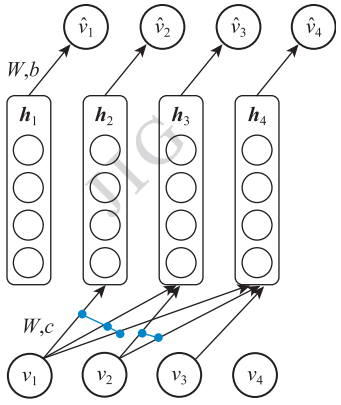


图9 神经自回归分布估计器示意图

Fig. 9 The illustration of neural autogressive distribution estimator

4 在目标和行为识别中的应用

由于深度网络可以无监督地从数据中学习到特征,而这种学习方式也符合人类感知世界的机理,因此当训练样本足够多的时候通过深度网络学习到的特征往往具有一定的语义特征,并且更适合目标和行为的识别。

4.1 谷歌的虚拟人脑

虚拟人脑是谷歌 2012 年研发出来的基于深度学习的具有自动学习能力的人工智能项目。它采用 1 000 台计算机共 16 000 个计算节点，利用 Youtube 网站上的视频作为训练集，花费 3 天的时间训练出 9 层的深度自编码器网络。其训练出来的深度神经网络已经可以模拟一些人脑的功能。例如，在完全没有标签的情况下，该网络能够自动地从训练集中学习到某些概念，例如，当输入是“猫”的图像时某些节点的响应会很强烈，而当输入的是“人脸”时另外一些节点会响应强烈，而且这些节点对于输入图像的旋转，平移等变化具有一定的不变性。

这些结果表明，仅仅通过无标签的样本来训练出某一类别的分类器是可行的，同时也说明了通过训练深度网络是可以让机器具有人一样的自主学习能力。这是人工智能领域的一个里程碑，其相关的研究成果[30]发表在国际机器学习大会（ICML2012）上，引起了世界范围内人们的关注。

4.2 大规模目标识别

2010 年开始的大规模视觉识别比赛（ILSVRC）是在 ImageNet 数据库[17]上进行的有关视觉目标识别的比赛。ImageNet 有超过 10 000 000 幅图像，超过 1 000 个类别，是目前公开的最大的视觉数据库，因此在这个数据库上进行的比赛反映了目前计算机识别技术的最高水平。前几年该项比赛最好的算法都是基于词袋模型或者可变部件模型，但是 2012 年基于深度学习的模型取得了最好的效果。他们在原始的 RGB 像素空间训练了深度卷积神经网络模型，该模型是有 6 000 万个参数，65 万个神经元构成的 5 层卷积网络。为了提高计算效率，他们采用 GPU 进行加速。该模型在识别和检测两个参赛单元上都超出了第 2 名至少 10 个百分点，这再一次说明了深度学习强大的学习能力。

4.3 图像的同时分类和标注

图像的分类和标注是计算机视觉中的两个重要问题。图像分类指的是对图像内容作整体的描述，例如给定一幅图像判断它属于“海滩”、“厨房”、“卧室”等预先定义好的类别中的哪一类；而图像的标注指的是对于图像中包含的内容作出判断，例如一幅图像中是否包含“天空”、“汽车”，“树木”等预先定义好的内容。很显然，这两个问题是相关的，虽然对于这两类问题，都有很多方法来解决各自的问题，但是却只有很少的工作尝试同时解决这两类问题[31]。对于这两类问题，目前最流行的方法是使用 LDA（latent dirichlet allocation）[32]来进行建模。但是 LDA 方法由于缺少闭式解而不得不使用变分近似方法或者马尔可夫蒙特卡洛方法，但是这些方法使用了过多的简化，而且计算速度慢。

给定一个类别，给出与之联系最紧密的标注单词和视觉单词。可以看出，SupDocNADE 可以学习到一定的语义信息：当给定类别“街道”的时候，与之联系最为紧密的标注单词是“建筑物”、“窗户”、“行人”和“天空”，而联系最密切的视觉单词都是类似于建筑物的表面或者窗户等。其他的类别也有相似的结果。

最近，Zheng 等人[33]提出使用基于神经自回归分布估计器（NADE）[29,34]的监督性神经自回归分布主题模型（SupDocNADE）来同时处理图像分类和标注问题。SupDocNADE 是一种基于 NADE 的监督性主题模型，其中每一个隐变量表示一个“主题”，而该隐变量的响应作为该“主题”的权重。SupDocNADE 使用“前向—反向”两步进行优化，在其优化过程中不存在像 LDA 一样的难于计算的归一化因子，因此整个模型可以准确地和有效地求解。文献[33]表明 SupDocNADE 在同时解决图像分类和标注问题中优于其他主题模型，并且会学习到具有语义特征的“主题”，如图 10 所示。

4.4 视频中的动作行为识别

正确快速地识别视频中人的动作行为对于视频搜索和视频监控具有十分重要的意义。目前识别视频中动作行为的技术基本都遵循以下几个基本步骤：首先检测时空显著兴趣点，接着在这些兴趣点的局部区域内提取特征描述符，然后对提取出来的特征点进行聚类形成字典，之后把这些特征进行最近邻量化并进行直方图向量汇总，最后利用 SVM 分类器对这些直方图特征向量进行分类训练和测试。最近几年深度学习技术被应用到视频里的动作行为识别中来。Ji 等人[35]提出多层的 3D 卷积神经网络来学习视频块的时空特征，并通过卷积操作来实现对整个视频的特征学习，从而替代之前的时空兴趣点检测和特征描述符提取。在 TRECVID 数据库上进行的实验取得了非常不错的效果。Baccouche 等人[36]提出使用稀疏卷积自编码器网络来学习视频块的时空特征，在 KTH 和 GEMEP-FERA 数据库上的实验结果表明其方法能与人工设计特征的方法取得类似的效果。Taylor 等人[37]提出使用卷积限制玻

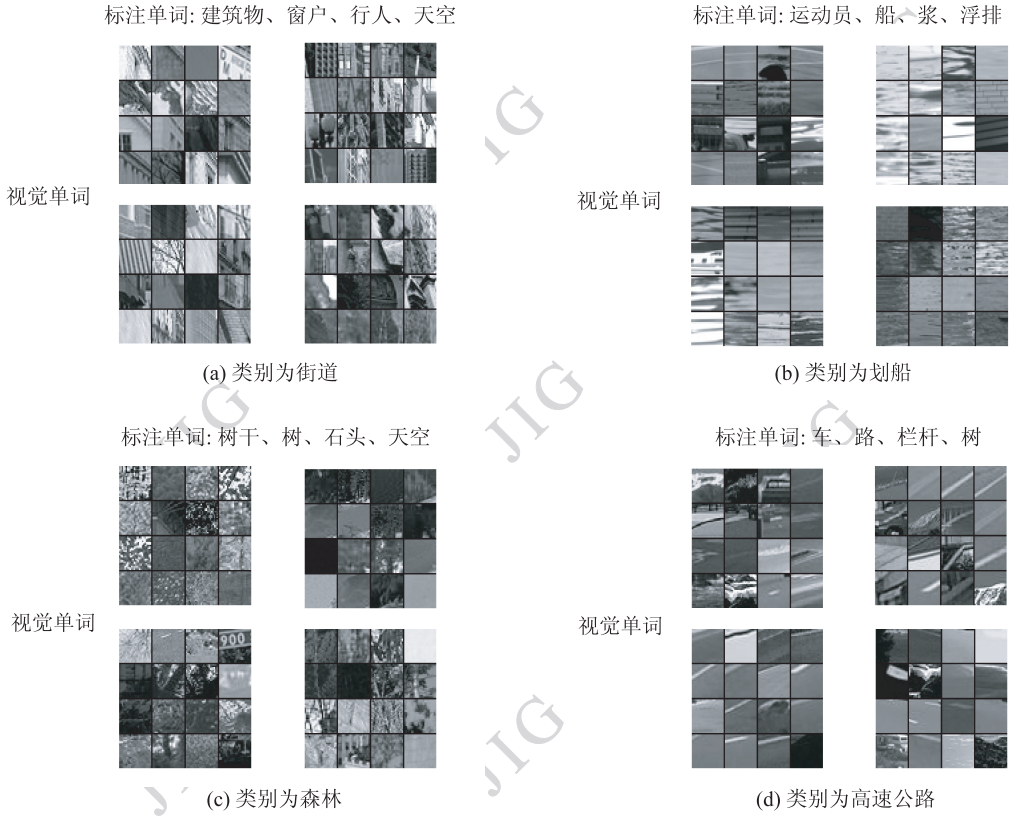


图 10　监督性神经自回归分布主题模型学习到的语义信息

Fig. 10　The semantic information learned from SupDocNADE

尔兹曼机[11]来学习视频中相邻两帧的时空特征，在 KTH 和 Hollywoods2 视频数据库中的对比实验结果表明，利用深度学习得到的特征与手工设计的 HOG、HOF 等特征具有类似的结果。

5　结　论

最后讨论一下深度学习与传统的多层神经网络之间的关系。笔者这里阐述一下自己的理解：

1）深度学习得到的是一个多层的深度结构，信号在这个多层结构中进行传播，最后得到信号的表达。这里的深度结构指的是层数多于一层的非线性函数关系，从这个角度来说，多层神经网络可以看做是深度学习的一个子类。当然，也存在其他结构的深度结构，例如文献[38-40]等。所以，从本质上说，任何一个多层的非线性结构都可以称之为深度结构，而深度学习则是学习这个深度结构的算法。本文介绍的两种主流深度学习算法可以理解成是对多层神经网络改进，但是不能就此将深度学习与深度神经网络等价，深度学习的概念内涵更加广。

2）深度学习的优势在于其能够学习到多层的非线性关系，这是其他的非深度学习算法所不能做到的。例如，决策树模型不能称之为一个深度网络，虽然它也是一个多层的结构，但是它的所有分支都可以通过枚举而变成一个一层的结构。这是由于深度学习的这种多层非线性特点使得深度学习能够用简洁的参数学习到复杂的函数关系。

3）深度学习不一定学习到的就是深度神经网络[38-40]，而神经网络也不一定是“深层”的（例如单层的神经网络）。同时，人们对于神经网络的定义也有狭义和广义之分，有些研究者甚至将一些类型的马尔可夫随机场模型也列入神经网络的范畴，而另一些研究者认为只有传统的前馈确定性的网络才可以称之为神经网络。

总之，深度学习的本质是学习到多层的非线性的函数关系。这种多层的非线性的函数关系使得人们能够更好地对视觉信息进行建模，从而更好地理解图像和视频。深度学习的优点在于模型的表达能力强，能够更好地处理诸如目标和行为识别这种非常复杂的问题，学习到更加复杂的函数关系，同时这

种方法也有一定的生物学基础。所以，可以期待未来会有更多的深度学习的结构单元和深度学习算法出现，来更好地解决这一问题。当然，深度学习也有一定的缺点，例如训练模型的时间比较长，需要不断地迭代来进行模型优化，不能保证得到全局最优解等等，这些也需要在未来不断地探索。目前深度学习的理论方面需要解决的主要问题有：1）深度学习的理论极限在那里，是否存在某个固定的层数，达到这个层数之后计算机可以真正实现“人工智能”；2）如何决定某类问题深度学习的层数和隐层节点的个数；3）如何评价深度学习得到的特征的优劣；4）对于梯度下降法如何进行改进以达到更好的局部极值点甚至是全局最优点。当前，在计算机视觉技术中对深度学习的研究和应用还比较少，因此对该领域的研究还有很大的研究空间。本文主要介绍了深度学习的原理以及目前在目标和行为识别中的研究进展，同时给出一些深度学习的应用成果，以希望能引起更多的人了解这一领域，并吸引更多的人在这一领域进行探索。

参考文献（References）

[1] Kong B. Comparison between human vision and computer vision [J]. Nature Magazine, 2002, 24(1): 51-55. [孔斌. 人类视觉与计算机视觉的比较 [J]. 自然杂志, 2002, 24(1): 51-55.]

[2] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]//Proceedings of IEEE Computer Society Conference on in Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005, 1: 886-893.

[4] Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions [C]//Proceedings of ICPR. Jerusalem, Irsael: IEEE, 1994, 1: 582-585.

[5] Matas J, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions [J]. Image and Vision Computing, 2004, 22(10): 761-767.

[6] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554.

[7] Hinton G E. Learning multiple layers of representation [J]. Trends In Cognitive Sciences, 2007, 11(10): 428-434.

[8] Hinton G E, Zemel R S. Autoencoders, minimum description length, and Helmholtz free energy [C]//Advances in Neural Information Processing Systems. Burlington, USA: Morgan Kaufmann, 1994: 3-10.

[9] Rumelhart D E, Hinton G E, Williams R J. Learning Representations by Back-Propagating Errors [M]. Cognitive Modeling: MIT Press, 2002, 1: 213.

[10] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders [C]//Proceedings of the 25th International Conference on Machine Learning. New York, NY, USA: ACM, 2008: 1096-1103.

[11] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations [C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: ACM, 2009: 609-616.

[12] Krizhevsky A, Hinton G E. Factored 3-way restricted boltzmann machines for modeling natural images [C]//Proceedings of International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: 2010: 621-628.

[13] Ranzato M A, Hinton G E. Modeling pixel means and covariances using factorized third-order boltzmann machines [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA: IEEE, 2010: 2551-2558.

[14] Lee T S, Mumford D, Romero R, et al. The role of the primary visual cortex in higher level vision [J]. Vision Research, 1998, 38(15-16): 2429-2454.

[15] Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex [J]. JOSA A, 2003, 20(7): 1434-1448.

[16] Arel I, Rose D C, Karnowski T P. Deep machine learning a new frontier in artificial intelligence research [J]. IEEE Transactions on Computational Intelligence Magazine, 2010, 5(4): 13-18.

[17] Deng J, Berg A, Satheesh S, et. al. Large scale visual recognition challenge [EB/OL]. [2013-11-14]. http://www.image-net.org/challenges/LSVRC/2012/index.

[18] Everingham M, Visual object classes challenge [EB/OL]. [2013-11-14]. http://pascallin.ecs.soton.ac.uk/challenges/VOC/.

[19] Hinton G E, Sejnowski T J, Ackley D H. Boltzmann Machines: Constraint Satisfaction Networks that Learn [M]. Pittsburgh, PA: Carnegie-Mellon University, Department of Computer Science, 1984.

[20] Andrieu C, de Freitas N, Doucet A, et al. An introduction to MCMC for machine learning [J]. Machine Learning, 2003, 50(1-2): 5-43.

[21] Hinton G E. Training products of experts by minimizing contrastive divergence [J]. Neural Computation, 2002, 14(8): 1771-1800.

[22] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks [C]//Advances in neural information processing systems. Vancouver, BC, Canada: MIT Press, 2007, 19: 153.

[23] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-

507.

[24] Hinton G E, Welling M, Teh Y W, et al. A new view of ICA [C]//Proceedings of Int. Conf. on Independent Component Analysis and Blind Source Separation. San Diego, CA: 2001: 746-751.

[25] Olshausen B A, Field D J. Sparse coding with an overcomplete basis set: a strategy employed by VI? [J]. Vision Research, 1997, 37(23): 3311-3326.

[26] Rehn M, Sommer F T. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields [J]. Journal of Computational Neuroscience, 2007, 22(2): 135-146.

[27] Lee H, Ekanadham C, Ng A. Sparse deep belief net model for visual area V2 [C]//Proceedings of Advances In Neural Information Processing Systems, Cambridge, MA: MIT Press, 2008: 873-880.

[28] Krizhevsky A, Nair V, Hinton G. The CIFAR-10 Dataset [DB/OL]. [2013-11-14]. http://www.cs.toronto.edu/~kriz/cifar.html.

[29] Larochelle H, Murray I. The neural autoregressive distribution estimator [C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, AISTATS 2011. Fort Lauderdale, FL, United states: Microtome Publishing, 2011: 29-37.

[30] Le Q V, Ramzato M, Monga R, et al. Building high-level features using large scale unsupervised learning, 1112.6209 [R]. New York, USA: Cornell University, 2012.

[31] Wang C, Blei D, Li F F. Simultaneous image classification and annotation [C]//Proceedings of IEEE Conference on in Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009: 1903-1910.

[32] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[33] Zheng Y, Zhang Y J, Larochelle H. A supervised neural autoregressive topic model for simultaneous image classification and annotation: 1305.5306 [R]. New York, USA: Cornell University, 2013.

[34] Larochelle H, Lauly S. A neural autoregressive topic model [C]//Advances in Neural Information Processing Systems. Nevada, United States: MIT Press, 2012: 2717-2725.

[35] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[36] Baccouche M, Mamalet F, Wolf C, et al. Spatio-temporal convolutional sparse auto-encoder for sequence classification [J]. Networks, 2005, 18(5-6): 602-610.

[37] Taylor G W, Fergus R, LeCun Y, et al. Convolutional learning of spatio-temporal features [C]//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010, 6316: 140-153.

[38] Ranzato M A, Huang F J, Boureau Y L, et al. Unsupervised learning of invariant feature hierarchies with applications to object recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN: IEEE, 2007: 1-8.

[39] Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional networks [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA: IEEE, 2010: 2528-2535.

[40] Ranzato M, Susskind J, Mnih V, et al. On deep generative models with applications to recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2011: 2857-2864.